Année universitaire 2025/2026

Module : analyse de données

Licence d'excellence (S5)

Travaux Pratiques (TP)

Sujet : Exploration du Dataset COVID-19

Pré-requis : Python, Pandas, Matplotlib, Seaborn, Scipy.
Scénario : Vous êtes Data Scientist et recevez un jeu de données cliniques brutes. Votre but est de déterminer si les tests sanguins permettent de prédire la positivité au COVID-19.

Exercice 1 : Initialisation et Analyse de Forme

1. Chargez le fichier dataset.xlsx.
2. Créez une copie du dataset nommée df.
3. Affichez les 5 premières lignes, le shape et les types de données (dtypes.value_counts()).

Exercice 2 : Gestion des valeurs manquantes (NaN)

1. Affichez une *Heatmap* des valeurs manquantes avec Seaborn (sns.heatmap(df.isna())). Que constatez-vous sur la structure des données ?
2. Calculez le pourcentage de NaN par colonne.
3. **Nettoyage :** Supprimez toutes les colonnes ayant plus de 90% de valeurs manquantes. Supprimez également la colonne Patient ID.
4. Vérifiez le nouveau shape.

Exercice 3 : Analyse de la Target

1. La target est SARS-Cov-2 exam result. Affichez la répartition des valeurs (positive/negative) en pourcentage (value_counts(normalize=True)).
2. Est-ce un dataset équilibré ? Quelle métrique de performance devriez-vous privilégier plus tard ?

Exercice 4 : Analyse des variables (Histogrammes)

1. Isolez les colonnes de type float (variables sanguines) et object (variables virales).
2. Créez une boucle pour afficher la distribution (sns.distplot) de chaque variable sanguine.
    - *Observation :* Les données sont-elles standardisées ?
3. Visualisez la relation entre les variables sanguines et la Target. Utilisez sns.distplot en colorant selon le résultat du test Covid (positif/négatif).
    - *Question :* Quelles variables semblent montrer une séparation entre les cas positifs et négatifs ? (Indice : Plaquettes, Leucocytes...).

Exercice 5 : Variables Catégorielles et Matrice de confusion

1. Utilisez pd.crosstab pour visualiser la relation entre la Target et les autres virus (ex: Influenza A).
2. Y a-t-il beaucoup de co-infections ?
3. Créez une nouvelle colonne est_malade qui vaut "vrai" si le patient a un autre virus que le Covid, et comparez cela avec la Target.

Exercice 6 : Corrélations et Tests Statistiques

1. Affichez la matrice de corrélation des variables sanguines (sns.heatmap(df.corr())).
    - Quelles variables sont très corrélées ?
2. **Test de Student :** Effectuez un ttest_ind (via scipy.stats) pour comparer les moyennes des taux sanguins entre les patients positifs et négatifs.
3. Affichez les variables pour lesquelles l'hypothèse nulle est rejetée (p-value < 0.05). Cela confirme-t-il vos observations visuelles de l'exercice 4 ?